Optimal Artificial Neural Network for the Diagnosis of Chagas Disease using Approximate Entropy and Data Augmentation

Maria Fernanda Rodriguez[†], Diego Rodrigo Cornejo[†], Luz Alexandra Díaz[†], Antonio G. Ravelo-García[§⋆], Esteban Alvarez[‡], Victor Cabrera-Caso[†], Dante Condori-Merma[†], Miguel Vizcardo Cornejo[†]

[†]Escuela Profesional de Física, Universidad Nacional de San Agustín de Arequipa, Perú

[§]Institute for Technological Development and Innovation in Communications, Universidad de Las Palmas de Gran Canaria, Spain

[⋆] Interactive Technologies Institute (ITI/LARSyS and ARDITI), 9020-105 Funchal, Portugal

[‡]Escuela de Física, Universidad Central de Venezuela, Venezuela

Abstract

The use of machine learning for disease diagnosis is gaining popularity due to its ability to process data and provide accurate results; however, its optimization remains a challenge. In the case of Chagas disease, endemic in Latin America and which has emerged as a health problem in more urban areas, early and accurate diagnosis is essential to prevent cardiac complications, since an estimated 65 million people are at risk of contracting it. This study used a database of 292 subjects distributed into three groups: healthy volunteers (Control group), asymptomatic Chagasic patients (CH1 group) and seropositive Chagasic patients with incipient heart disease (CH2 group). A densely connected neural network was used to classify them into the group to which they belonged. The network received as input the Approximate Entropy values of each individual, which were calculated from the 24-hour circadian profiles every 5 minutes (288 RR subsegments). In addition, time series data augmentation algorithms were applied during the training phase to improve the classification results. This approach allowed to reach 100% accuracy and precision, validated by the ROC curve with AUC values of 1. Thus, the efficiency demonstrated by the neural network suggests that increasing the amount of training data may be crucial to optimize the early diagnosis of cardiac involvement that may develop Chagas disease, and consequently, it could be a determining factor in refining machine learning in this area.

1. Introduction

Chagas disease, or American trypanosomiasis, is caused by Trypanosoma cruzi. This vector is present in 21 continental countries in the Region of the Americas and about 65 million people are at risk of contracting the infection, which causes approximately 12,000 deaths annually [1], and in recent decades it has begun to be detected in other non-endemic regions of the Americas [2]. The disease presents acutely and, if not diagnosed and treated in time, becomes a chronic disease. The most important consequence is chronic chagasic cardiomyopathy, which occurs in 20-40% of infected persons and can be potentially lethal. As it is considered a neglected tropical disease [2], the use and optimization of non-invasive and low-cost diagnostic tools is paramount. In this context, machine learning has become popular as a promising technique for disease diagnosis, including Chagas disease [9–12]. Despite this, optimization of these tools remains a challenge, due to, among other reasons, the limited amount of available data, highlighting the need to implement data augmentation techniques to improve their efficiency.

Although image analysis techniques are widely used, heart rate variability (HRV) analysis can be very useful due to its prognostic significance. In particular, Approximate Entropy has been shown to be a valuable statistic for the study of congestive heart failure [5, 6], one of the main clinical manifestations of Chagas disease [7], and it was also used to identify significant differences at different times of the day between groups of patients with this disease [8]. This is why the present work proposes the development and optimization of a densely conected neural network using the HRV based on the approximate entropy of a database of patients with Chagas disease, using data augmentation techniques, which, although are more common in image analysis, they are also applicable to time series.

2. Method

2.1. Database

This study made use of the database of the Tropical Medicine Institute of the Universidad Central de Venezuela, which includes information on 292 individuals who underwent various tests with their respective informed consent. These tests included clinical evaluation, Gerreiro Machado-Serology test, chest X-ray, echocardiogram, electrocardiogram and Holter recording (24 hours). The patients and volunteers were divided into three groups: the Control group, consisting of 83 healthy persons (volunteers), the CH1 group composed of 102 infected patients only with positive Machado-Gerreiro serology test, and the CH2 group composed of 107 seropositive patients with incipient heart disease, involvement of first-degree atrioventricular block, sinus bradycardia or right bundle branch block of His and were not receiving treatment or medication. ECG signals were recorded at a frequency of 500 Hz with a resolution of 12 bits.

2.2. Data preprocessing

Obtaining the QRS complexes from the ECG was performed with the Pan-Tompkins [13] algorithm, then generating the 288 5-minute RR tachograms for each subject from the database. In addition, a filter used in [8] was implemented to remove noise.
Considering that we are working with time series data, Approximate Entropy (ApEn) was applied to each 5-minute RR subsegment of each subject according to the definition given by Pincus [14], where if the time series data consists of N elements:

$$ApEn(m, r, N) = -\frac{1}{N-m} \sum_{i=1}^{N-m} \log\left(\frac{A_i}{B_i}\right) \quad (1)$$

where m is the embedding dimension, r is a threshold and A_i and B_i are the proximity measures between the embedding vectors in m and $m+1$ dimensions respectively.

After testing values of m from 1 to 4, and r from 10 to 50 standard deviation, the parameters $m = 2$ and $r = 40\%$ of standard deviation were finally selected because of the good discrimination they achieved among the three groups and each group with another.

Finally, some missing ApEn data (produced by noise filtering and the database itself) were interpolated using the Matlab function fillgaps in order to predict missing data in a series. Thus, each subject was characterized by a complete record of 288 ApEn values.

2.3. Network architecture and data augmentation

For the purpose of this work, first, all data were randomly divided as follows: 70% of 292 subjects formed the training set and the other 30% the test set. In addition, a validation set was considered and involved 30% of the training set during the network training phase.

A Densely Connected Neural Network was implemented in Pyhton, using the Keras and Scikit-learn environments, with a sequential model and dense layers. 288 ApEn values were the input layer nodes, which were previously standardized. The outputs correspond to the three groups in which they are located: Control, CH1 and CH2.

The Adam optimizer was used with a small learning rate, the loss function was categorical cross entropy, and the activation function is chosen according to the training, except in the last layer, where it was softmax. All other hyperparameters are selected according to the evolution of the network training too.

To increase the performance of the model, data augmentation techniques were proposed, because of their ability to increase the generalization capacity of machine learning models, to increase the samples (subjects) of the training set. Most of these are inspired by image recognition. Thus, scaling and jittering algorithms were selected as augmentation algorithms because of their great ability to preserve the temporal pattern of the data [15].

3. Results

An optimal 3-hidden layer architecture was found with 15, 10 and 8 neurons respectively. The activation function was sigmoid in all layers except the output layer. The Adam optimizer was used with a learning rate of 0.002, 200 was the epoch limit and the batch size was 10. Also, overfitting was controlled thanks to the earlystopping function implemented to stop the model training when the validation loss does not reach lower values in 5 consecutive epochs.

To visualize the performance of the model, first, without having implemented data augmentation, Figure 1 shows the confusion matrix. With it, the classification results of our model were as follows: for the Control group we obtained an accuracy of 0.889, recall of 0.960 and F1-score of 0.923. For the CH1 group, the accuracy was 1.000, recall 1.000 and F1-score 1.000. And for the CH2 group the results were 0.969 for precision, 0.912 for recall and 0.939 for F1-score. The accuracy of the model was 95.5%, and the overall weighted precision was 95.6%.
To observe the success rate, the receiver operating characteristic curve (ROC curve) was plotted. being a multi-class classification, an extended version of the ROC curve had to be applied with the micro and macro averaging algorithm

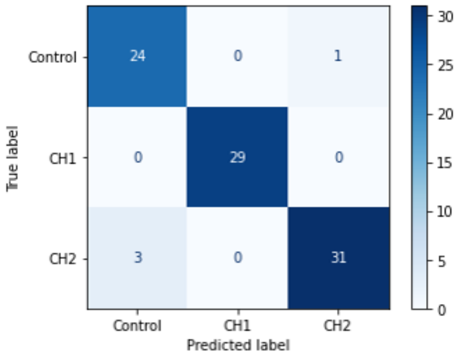

Figure 1. Confusion matrix without data augmentation

in the scikit-learn library.

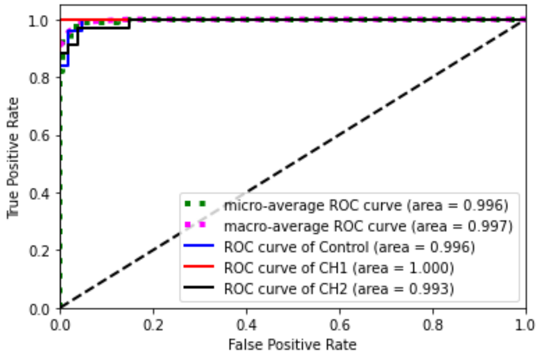

Figure 2. ROC curve without data augmentation

Thus, Figure 2 shows one curve for each group (one versus all) and two general curves for the entire classification. As all AUC values are very close to 1, a good model performance is confirmed despite the fact that no regularization or data augmentation technique was used; however, as these are neural networks, the percentage can be improved.

By applying the data augmentation techniques mentioned above, but keeping the same network architecture, rows (patients) were added to the standardized ApEn matrix of the original training set. In this sense, the network was trained with 3 times the size of the original training set (three times the number of patients). Data augmentation was not applied to the test set to evaluate the performance of the network with original data.

The confusion matrix, using the same data split as above (70% training and 30% test) is plotted in Figure 3. From it it can be seen that all evaluation metrics (precision, recall and F1-score) for each group and overall, were unity, thus having 100% overall model accuracy and precision, demonstrating an excellent classification result. This is validated by the multiclass ROC curve (Figure 4), whose AUC values were exactly unity.

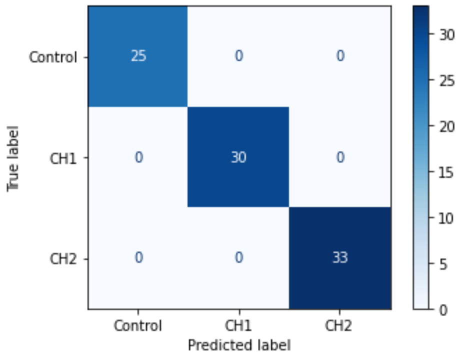

Figure 3. Confusion matrix with data augmentation

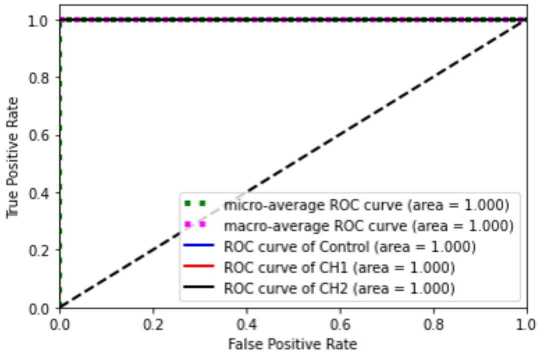

Figure 4. ROC curve with data augmentation

Also, taking into account that with the increase of data the number of patients with which the network trains is numerous, it was possible to vary the division of the original data sets (training and test), whose general classification metrics are summarized in the table 1, finding that the accuracy of the model is higher than 90% even when only 30% of the original patients are used for training.

Table 1. Overall evaluation metrics, using data augmentation for different original patient divisions

Test set	**Accuracy**	**Overall precision**	**Weighted overall precision**
30% (88 subjects)	100.0%	100.0%	100.0%
40% (117 subjects)	98.3%	98.1%	98.3%
50% (146 subjects)	97.3%	97.1%	97.3%
60% (176 subjects)	94.9%	94.8%	95.0%
70% (205 subjects)	90.7%	90.5%	90.7%

4. Discussion and conclusions

Approximate Entropy proved to be a powerful statistical tool to characterize and discriminate time series, since, considering the differences in the circadian profiles obtained through its use among the three groups of the database, by implementing a deep neural network to classify them, a good performance of the model was obtained without using regularization techniques or data augmentation algorithms. Accuracy values higher than 88% were achieved for all groups, according to all the evaluation metrics considered, as well as an accuracy of 95.5% and an overall precision of 95.6%. However, despite obtaining comparable results to those of previous research based on clinical and sociodemographic data for the prediction of Chagas disease [11], we chose to implement data augmentation algorithms because of the importance of diagnostic optimization for obtaining reliable and accurate results.

With the data augmentation, scaling and jittering algorithms, it was possible to enrich the training data set and improve the predictive capacity of the model, obtaining an excellent classification capacity: 100% accuracy and precision for the same division of the original data. Likewise, by training with three times as many patients from the training set, it was possible to decrease the number of original patients used for training, obtaining a classification accuracy of more than 90% even when only 30% of patients from the original database were used. This result clearly reflects the relevance of data augmentation in neural networks when working with databases that present a limited number of samples, and at the same time, supports the use of this technique in time series analysis.

Early and non-invasive identification of patients with Chagas disease who are not yet symptomatic (CH1 group) is crucial for successful treatment and improvement in the quality of life of patients. In this regard, the proposed approach is presented as a highly reliable and low-cost diagnostic tool. The application of this approach could be a promising alternative to improve medical care and disease management, which can have a significant impact on the health of the general population.

Acknowledgements

Universidad Nacional de San Agustin de Arequipa

References

[1] Pan American Health Organization (2021). *Enfermedad de Chagas e Inmunosupresión. Decálogo para la prevención, el diagnóstico y el tratamiento.* https://iris.paho.org/handle/10665.2/54561

[2] World Health Organization(2010). First WHO report on neglected tropical diseases: working to overcome the global impact of neglected tropical diseases. In *First WHO report on neglected tropical diseases: Working to overcome the global impact of neglected tropical diseases* (pp. 172-172).

[3] Marin-Neto, J., Cunha-Neto, E., Maciel, B. & Simões, M. (2007). Pathogenesis of chronic Chagas heart disease. *Circulation, 115*(9),1109-1123. doi:10.1161/CIRCULATIONAHA.106.624296.

[4] Di Lorenzo Oliveira, C., Nunes, M. C., Colosimo, E., ... & Ribeiro, A. L. P. (2020). Risk Score for Predicting 2-Year Mortality in Patients With Chagas Cardiomyopathy From Endemic Areas: SaMi-Trop Cohort Study. *Journal of the American Heart Association, 9*(6), e014176. doi: 10.1161/JAHA.119.014176.

[5] Beckers, F., Ramaekers, D. & Auber, E. (2001). Approximate Entropy of Heart Rate Variability: Validation of Methods and Application in Heart Failure. *Cardiovascular Engineering: An International Journal, 1*, 177–182. https://doi.org/10.1023/A:1015212328405

[6] Namazi, H., Baleanu, D. & Krejcar, O. (2021). Age-Based Analysis of Heart Rate Variability (hrv) for Patients with Congestive Heart Failure. *Fractals, 29* (3), 2150135-1073. doi:10.1142/S0218348X21501358

[7] Rassi, AJr., Rassi, A. & Marin-Neto, JA. (2009). Chagas heart disease: pathophysiologic mechanisms, prognostic factors and risk stratification. *Mem Inst Oswaldo Cruz, 1*, 152-158. doi:10.1590/s0074-02762009000900021

[8] Vizcardo, M., & Ravelo, A. (2018). Use of Approximation Entropy for Stratification of Risk in Patients With Chagas Disease. In *2018 Computing in Cardiology Conference (CinC)*, 45, 1-4. doi:10.22489/CinC.2018.234

[9] Cochero, J., Pattori, L., Balsalobre, A., Ceccarelli, S. & Marti, G. (2022). A convolutional neural network to recognize Chagas disease vectors using mobile phone images. *Ecological Informatics, 68*, 101587. https://doi.org/10.1016/j.ecoinf.2022.101587.

[10] Sanchez-Patiño, N., Toriz-Vazquez, A., Hevia-Montiel, N. & Perez-Gonzalez, J. (2021). Convolutional Neural Networks for Chagas Parasite Detection in Histopathological Images. *International Conference of the IEEE Engineering in Medicine & Biology Society*, 2732-2735. doi:10.1109/EMBC46164.2021.9629563.

[11] De Santana, W., Machado, A., Júnior, P., de Melo, C., ... & Jeraldo, V. (2021). Machine learning and automatic selection of attributes for the identification of Chagas disease from clinical and sociodemographic data. *Research, Society and Development, 10*(4), e19310413879-e19310413879. https://doi.org/10.33448/rsd-v10i4.13879

[12] Pereira, A., Mazza, L., Pinto, ... & Soares, G. (2022). Deep convolutional neural network applied to Trypanosoma cruzi detection in blood samples. *International Journal of Bio-Inspired Computation, 19*(1), 1-17. doi:10.1504/IJBIC.2022.10044882

[13] Pan, J., & Tompkins, W. J. (1985). A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering, 32*(3), 230–236. doi:10.1109/TBME.1985.325532

[14] Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences, 88*(6), 2297–2301. doi:10.1073/pnas.88.6.2297

[15] Iwana, B. & Uchida, S. (2020). Time Series Data Augmentation for Neural Networks by Time Warping with a Discriminative Teacher. *International Conference on Pattern Recognition*, 3558-3565. 10.1109/ICPR48806.2021.9412812